Case studies in Gaussian process modelling of computer codes

Marc C. Kennedy   Clive W. Anderson   Stefano Conti *
Anthony O'Hagan

Department of Probability and Statistics
Centre for Terrestrial Carbon Dynamics
University of Sheffield

Abstract

Introduction

Broad comprehension of complicated physical processes is increasingly often developed and supported by means of sophisticated mathematical models implemented within computer codes. It is, however, well known that before relying upon the explanatory and predictive abilities of any computer simulation, a variety of validatory checks need to be carried out.

In the context of computer experimentation, the practical complications casting most serious doubts on how adequately and realistically a computer model reproduces reality usually arise from: vague or controversial believes about the value of some of the code's parameters; availability of limited and/or inaccurate driving data; restrictions due to the CPU cost required for actually running the program; incomplete or imperfect knowledge of the real-world phenomenon of interest. In order to identify and attenuate the main sources of uncertainty hampering a program's performance several statistical methods have already been proposed in the classical literature (see Saltelli et al., 2000a, for an exhaustive reference).

The Bayesian perspective

Over the past decade interesting results were obtained from addressing problems related to computer model uncertainty in a Bayesian fashion. In

*Department of Probability and Statistics, The Hicks Building, University of Sheffield, Sheffield S3 7RH, UK. Tel: +44(0)114 222 3824; fax +44(0)114 222 3759; e-mail: s.conti@sheffield.ac.uk

particular, a convenient and flexible strategy consists in assigning a semi-parametric Gaussian process prior to the program's response; details of the technique can be found e.g. in Oakley and O'Hagan (2002). Preliminary *emulation* of a code by such means has already been fruitfully exercised on simulators of nuclear radiation releases (Kennedy and O'Hagan, 2001) and on models for vehicle crash and spot welding (Bayarri et al., 2002). Besides relevant specific findings, results from these case-studies emphasise how widely applicable and enlightening the principle of Gaussian process-based emulation can be.

CTCD

The Centre for Terrestrial Carbon Dynamics (*CTCD*) is a consortium of British academic and governmental institutions, established for the purpose of progressing the scientific understanding of the role played by terrestrial ecosystems in the carbon cycle, with stress on forest ecosystems. CTCD is funded by the Natural Environment Research Council for 5 years as one of its national centres of excellence in earth observation. The ultimate goals of the project are: to gauge carbon fluxes and their uncertainties at different space/time resolutions; to devise methodological, data and instrument advances for reducing these uncertainties; to deliver relevant findings in accessible formats to the scientific community and to policy makers. These tasks are pursued with the support of a variety of environmental models designed for simulating carbon patterns over different geographical and climatic scenarios. Unfortunately, such models suffer from coarse reproduction of some underlying physical processes and loose connections to driving data.

Within the Centre, Bayesian methods are being employed for the assessment of the relevant model (and data) developments required for reducing the uncertainty around them. In this setting, statistical challenges presently requiring special care appear to be:

prediction: estimation of (possibly functionals of) model outputs at input configurations other than the available ones;

uncertainty analysis: exploration of the output distribution induced by assigning some probability distribution to uncertain inputs;

screening: identification of which of the code inputs are significantly *active*, i.e. most influential on the outputs;

sensitivity analysis: examination of how model outputs react to changes in appropriate inputs;

code verification: detection of bugs in the actual implementation of the program.

We will present three case studies of the Bayesian approach addressing these challenges in the context of a local (*SPA*: Williams et al., 1996), a regional (*ACM*, a simpler upscaled version of SPA: Williams et al., 1997) and a global (*SDGVMd*: Lomas et al., 2002) vegetation model.

Case Study 1 - A variety of extensions and improvements to SDGVMd were undertaken in the first year of CTCD's operation. Simple sensitivity analysis exercises served to identify problems with the evolving code. Yearly averages were computed over 100 years for the principal model output (Net Ecosystem Productivity, or *NEP*). Sensitivity was explored by examining a few inputs at a time. In each analysis, uniform probability distributions were assigned to 5 relevant soil and plant inputs, while the remainder were fixed at suggested default values. Plots of main effects (Saltelli et al., 2000b) proved a cheap and effective confirmatory tool by uncovering faults in one of the coded algorithms and recognising which of the considered inputs NEP is significantly sensitive to.

Case Study 2 - Motivated by an earlier investigation (Williams et al., 2001), a variety of analyses have been performed on ACM and SPA. The simplest of these consisted just in replicating the uncertainty analysis in Williams et al. (2001) using Gaussian process emulation. As with other reported applications of the Bayesian approach, we were able to accurately replicate the findings of the earlier Monte Carlo analysis using a small fraction of the code evaluations in the MC sample. The finding that Leaf Area Index was the most active input, was augmented and interpreted by plotting main effects.

A more complex analysis arises from the recognition that ACM is a kind of emulator of SPA, designed to operate on a larger geographical scale and when values for some of SPA's inputs may not be available. We expect to meet extrapolation problems as well, when applying the more global scale SDGVMd outside the relatively data-rich region of Northern Europe. It was therefore a useful exercise to employ Gaussian process emulation to provide an alternative simplification of SPA, in the context for which ACM was designed. Our emulator, based on only 150 runs of SPA, outperforms ACM (which had been built using 6561 SPA runs).

Case Study 3 - A major deliverable of CTCD will be an estimate of the UK carbon budget, in April 2004, using SDGVMd. A preliminary version of this estimate is to be given at a "UK Carbon meeting" scheduled in Sheffield on January 2004. We will quantify uncertainty on the UK carbon budget using Bayesian methods, recognising uncertainty in major model parameters defining vegetation and soil properties. In this talk, we will describe the methodology used to derive this uncertainty/sensitivity analysis by aggregation of pixels covering the whole UK.

In conclusion, the proposed Bayesian approach to computer experimentation has already supplied useful insights to CTCD modellers and is expected to yield profitable responses when applied to more demanding test beds. Uncertainty and sensitivity analyses will be integral parts of all major CTCD deliverables.

References

M. J. Bayarri, J. O. Berger, D. Higdon, M. C. Kennedy, A. Kottas, R. Paulo, J. Sacks, J. A. Cafeo, J. Cavendish, C. H. Lin, and J. Tu. A framework for validation of computer models. In D. Pace and S. Stevenson, editors, *Foundations for Verifications and Validation in the 21st Century*, 2002.

M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 63(3):425–464, 2001.

M. R. Lomas, F. I. Woodward, and S. Quegan. The role of dynamic vegetation models. Technical report, University of Sheffield, Sheffield UK, 2002.

J. E. Oakley and A. O'Hagan. Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika*, 89(4):769–784, 2002.

A. Saltelli, K. Chan, and E. M. Scott, editors. *Sensitivity analysis.* Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 2000a.

A. Saltelli, S. Tarantola, and F. Campolongo. Sensitivity analysis as an ingredient of modeling. *Statist. Sci.*, 15(4):377–395, 2000b.

M. Williams, E. B. Rastetter, D. N. Fernandes, M. L. Goulden, S. C. Wofsy, G. R. Shaver, J. M. Melillo, J. W. Munger, S. M. Fan, and K. J. Nadelhoffer. Modelling the soil–plant–atmosphere continuum in a *Quercus–Acer* stand at harvard forest: the regulation of stomatal conductance by light, nitrogen and soil/plant hydraulic properties. *Plant, Cell and Environment*, 19:911–927, 1996.

M. Williams, E. B. Rastetter, D. N. Fernendes, M. L. Goulden, G. R. Shaver, and L. C. Johnson. Predicting gross primary productivity in terrestrial ecosystems. *Ecological Applications*, 7:882–894, 1997.

M. Williams, E. B. Rastetter, G. R. Shaver, J. E. Hobbie, E. Carpino, and B. L. Kwiatkowski. Primary production of an arctic watershed: an uncertainty analysis. *Ecological Applications*, 11(6):1800–1816, 2001.